ASSIGNING PUTATIVE GENE FUNCTIONS TO MAPPED PROBE LOCI IN THE GRAINGENES GENOME DATABASE AND SEQUENCING OF WHEAT ENDOSPERM cDNAs

Gerard R. Lazo[1], Lance A. Larka[1], Cheryl C. Hsia[1], Kent F. McCue[1], Mark E. Sorrells[2], David E. Matthews[2], Melinda Au[3], Nancy A. Federspiel[3], and Olin D. Anderson[1]

1. USDA ARS, Western Regional Research Center, Albany, CA 94710-1105
2. Department of Plant Breeding & Biometry, Cornell University, Ithaca, NY 14853
3. Stanford DNA Sequence and Technology Center, Palo Alto, CA 94304

Abstract

The genome database GrainGenes has been a useful tool for collecting and displaying research information for the cereals community. A major component of the data in GrainGenes is genetic maps and the associated information on probes used to construct those maps. In many instances the probes were derived from random cDNAs. Although these cDNAs were used as probes to map specific genes, nothing was known of their identity and/or function. Efforts were taken to gain sequence data from these mapped probes. In addition, random cDNAs from a wheat endosperm library were single-pass sequenced to evaluate the usefulness of such data and to assess some of the parameters that would be involved in a larger-scale expressed sequence tag (EST) program. In this effort sequences were compared against NCBI genome data banks using BLAST search programs. To date, over 1000 sequences have been analyzed. About 26% of the mapped sequences and 49% of the wheat endosperm library sequences had good matches to archived database sequences. Many of the mapped sequences were unique. Some of the endosperm sequences were represented more than once, and were associated with storage proteins and common housekeeping enzymes. Sequences which did not match wheat sequences in the database often matched sequences of closely related cereals. This data is being integrated into the GrainGenes database and display tools are being developed to make searches within the database more informative.

Sequencing

From the GrainGenes probe repository, 576 clones were selected for sequencing. These were from a collection of CDO (oat) and BCD (barley) cDNA clones which have been used for mapping purposes within the Gramineae (Heun, Kennedy, Anderson, Lapitan, Sorrells, Tanksley. 1991. Genome 34:437-447). The remaining clones sequenced (about 300) were derived from a cDNA library constructed from wheat embryos over a time-course after anthesis. EcoRI-linker adapters were added to mRNA and cloned into lambdaZAP II, and the released phagemid was used for sequencing. The collection of cDNAs were single-pass sequenced on ABI 377 (Stanford DNA Sequencing and Technology Center) and ABI 310 (USDA ARS WRRC) machines. Trace files and results were tranferred to a UNIX environment for further processing.

Data Analysis

The ACEDB (Durbin and Thierry-Mieg. 1991-. ACEDB. ftp://ftp.sanger.ac.uk/pub/acedb/) database program was used to manage the sequence information for the GrainGenes genome database. ACEDB

has a long history in genome research and will run on a wide variety of computer platforms (UNIX, Win95/NT, MacOS, Java, WWW). Each version presents information in textual- and graphical-user interfaces and allows searches with it's powerful query language. The GrainGenes database can be accessed over the World-Wide-Web at http://probe.nalusda.gov/ or from the central WWW site located at http://wheat.pw.usda.gov/. The cDNA sequences were visualized using the Genetic Data Environment (GDE) program (Smith, Overbeek, Woese, Gilbert and Gillevet. 1994. CABIOS 10:671-675) to remove vector-contaminating sequences (shown below), and the insert DNA sequences were submitted to the NCBI BLAST (Altschul, Gish, Miller, Myers, and Lipman. 1990. Mol. Biol. 215, 403-10) mail server, comparing sequences to the non-redundant (nr) and expressed sequence tag (est) Genbank databases (v.104.0, 316,258 and 1,364,418 sequences, respectively. Retrieved BLASTN reports were converted to an ACEDB format using Perl-programmed scripts. Within the ACEDB format, BLASTN alignments were grouped for GrainGenes database comparisons among the genomes for Wheat, Barley, Oat, Rye, Corn, Rice, Sugarcane, Other Grass Species, and Non-Grass Species. Within the database, using the ACEDB Sequence-class graphical display, BLASTN alignments could be compared to query sequences using color-codes to distinguish genome and database origins (see color index). A different Method-class was established for each genome of interest.

Sequence Display

The ACEDB Sequence display (A) shows colored BLAST scoring boxes. The width is determined by the High-scoring Segment Pair (HSP) score from BLASTN. A score of 400 (arbitrary cutoff score) or above yields a box of maximum width. The coded colors are ranked based on high score and several alignments may be stacked within a single Method-class. Methods-classes are easily configured. By "clicking" a color-coded box the region annotated by the BLASTN alignment is highlighted on the DNA sequence. A "double-click" will bring up more information about the aligned sequence within a text-display (B). Bold-faced text is a hyperlink to additional information. Under the heading "DNA_homol" may be hyperlinks to other sequences that also have similar BLASTN alignments with the shown sequences. The full spectrum of matches can be analyzed here. In some cases where there is no Genbank sequence data loaded, a WWW hyperlink derived from Genbank index files has been constructed which opens a sequence information page at the NCBI WWW site (C). By "clicking" the named query sequence (the cDNA sequence) in the graphical display, a text-display showing the database origins, aligned sequence names, and the scores and respective alignments is presented (D). Selecting the aligned sequence name will again open another text display.

Blixem Display

The sequences which have BLASTN alignment boxes shown may be further inspected to show all sequence alignments within the GrainGenes database using the integrated software tool, Blixem (Sonnhammer and Durbin. 1994. CABIOS 10:301-307). Blixem allows inspection of BLAST sequence alignments (A) and also allows dot-plot comparison (B) of any two sequences. An alignment window (C) shows the region with sequences shown in sense and reverse-compliment orientations.

Map Display

GrainGenes has a collection of genetic maps which cover chromosomes from many of the small grain cereals. These maps have been generated from several research laboratories on numerous germplasm, and in many cases, using cDNA probes used solely as genetic markers with no assigned function. Conditions have been set within the graphical ACEDB Map display to highlight loci (shown in yellow) for which probes exist that have been sequenced and have positive BLASTN results. One benefit of this display is that it can easily be used to screen maps of different species and may be useful in establishing

syntenic relationships among different maps maintained in the GrainGenes database. "Clicking" on a highlighted locus will lead to information about the probe and it's BLASTN results. Illustrated is a sample of candidate gene assignments that might be linked to loci within a single chromosome (6A) map of wheat.

Query Language

The ACEDB program also has a query- and table-building feature for constructing complex searches of information in the database. Shown are some sample queries with explanation of their function shown below:

- 1. Find all BLAST hits which have Sequence Titles loaded in the database (gene name search).
- 2. Find all sequenced probes which have BLASTN hits with an HSP score greater than 400 (high-scoring sequences).
- 3. Find titled BLAST nr/est hits which align to more than 10 sequences in the cDNA clone collection (abundant sequences).
- 4. Find sequences in the cDNA clone collection which do not have BLAST nr/est hits (unidentified sequences).
- 5. Find map loci defined by probes which map to more than three genetic maps (synteny mapping).

ACEDB Classes

In addition to the Map, Probe, and Sequence classes in GrainGenes other classes of biological relevance are available. Classes are selected from the Main Menu screen. Other information includes Genes (Wheat Gene Catalogue information including synonyms, alleles, and chromosomes), Germplasm (genotypes and pedigrees of cultivars, genetic stocks, and wild accessions), Pathology (descriptions of plant diseases of cereals including many full-color images), Colleague lists and References.

Summary

Sequence data has been obtained from at least a single pass for 824 individuals including probe repository clones and random wheat endosperm cDNAs. These sequences yielded 3064 independent sequence hits with BLAST HSP scores above 400 in the Genbank non-redundant (nr) database. This represents 18% and 40% of the Barley and Oat sequences, respectively. For the endosperm cDNAs, 32% had HSP scores greater than 400. The sequences identified by BLAST searches span many of the Gramineae as well as others outside the species. Many sequences which did not match the Genbank nr database, often matched sequences in the est database. Of the 539 genetic maps constructed in GrainGenes, probe sequences with positive BLASTN reports correspond to 2609 probe loci distributed on 250 maps. These included maps of wheat, barley, oat, rye, sugarcane, rice, and maize. The occurrence of common markers over several maps can be used to develop syntenic relationships. These can be displayed using the Multimap-class of ACEDB (not shown). Lists of all mapped probes and their putative functional assignment were created using ACEDB query language and table-maker functions. Sequence titles were assigned to various biochemical categories. Shown is a sample of BLASTN hits by function.

Candidate Genes Identified

STORAGE PROTEINS			

gliadin, alpha-/beta-/gamma-	LMW glutenin	HMW glutenin, Glu-1-2/Glu-A3/B3/D1-2b/Glu-D3	grain softness protein, GSP-1a/GSP-1b
puroindoline, a/b	endosperm lumenal binding protein	thionin, alpha-1/type V	alpha-2-purothionin.
globulin	avenin	hordein, gamma	gamma-secalin (related) storage protein
serpin	seed storage protein		
CELL MACHINERY			
rRNA, 12S/16S/18S/28S	ubiquitin	elongation factor, 1-alpha/1-beta	heat shock protein, 70/82/90
translation initiation factor	pathogenesis-related protein	vaculolar membrane	profilin protain kinase
protein kinase.	shaggy-like kinase	Barperm1 (perm1)	snRNP-related protein
victorin binding protein	Wilm's tumor suppressor		
CELL STRUCTURE			
tubulin, alpha-/beta-	extensin-like protein	histone, H2A/H2B/H3/H4	invertase
actin			
STORAGE METABOLISM			
UDP-glucose pyrophosphorylase	starch branching enzyme I	amylase, alpha	amylase, alpha inhibitor
CELLULAR METABOLISM/MAINTENANCE			
S-adenosylmethionine synthetase	S-adentlmethionine decarboxylase	glyceraldehyde-3-phosphate dehydrogenase	ribulose-1,5-bisphosphate carboxylase
5-bisphosphate carboxylase small subunit	ATP/ADP translocator	ATP synthase beta subunit	H+-pyrophosphorylase
vacuolar membrane proton-translocating	orthophosphate dikinase	transmembrane proton pump	chlorophyll a/b binding
ABA-induced protain	acyl-CoA synthetase	disulfide isomerase	betaine aldehyde dehydrogenase
beta-D-glucan exohydrolase, isoenzyme ExoII	delta-1-pyrroline-5-carboxylate synthase	glutathione reductase	glutathione S-transferase gene
alanine aminotransferase	glycine decarboxylase	cysteine proteinase	tryptophan synthase
peptidylprolyl cis-trans isomerase	phosphoglycerate kinase	glyoxysomal malate dehydrogenase	phenylalanine ammonia-lyase
3-hydroxy-3-methylglutaryl coenzyme A reductase	methyljasmonate-inducible lipoxygenase	permatin	ras-related GTP binding protein